内群体偏爱或黑羊效应? 经济博弈中公平规范执行的群体偏见*

张　振[1] 齐春辉[1] 王　洋[2] 赵　辉[1] 王小新[1] 高晓雷[3]

([1]河南师范大学心理学院, 新乡 453007) ([2]国网天津市电力公司东丽供电分公司, 天津 300300)
([3]西藏大学教育学院, 拉萨 850000)

摘　要　公平规范执行代指人们自愿损耗利益来惩罚违反公平原则行为的意愿和行为，被视为人类社会的重要特征及维持合作行为的关键因素。群体认同是指个体对其所属群体身份的知觉及其所付诸于该群体身份上的价值与情绪，直接影响着群际互动中人们的公平规范执行行为。基于多种资产分配任务，国内外学者发现公平规范执行中群体偏见存在两种矛盾的表现形式：人们有时更愿意接受内群体成员的不公提议，表现为内群体偏爱现象(the in-group favoritism, IGF)，有时又更愿意拒绝内群体成员的不公提议，表现出黑羊效应(the black sheep effect, BSE)。当前，纯粹偏好理论和规范聚焦理论常被用来解释上述两种相悖的现象。未来研究应侧重从多种视角揭示公平规范执行偏见的边界条件，比较多种线索操纵所致偏见的差异性，促进两种理论的融合与补充，并增强其潜在神经机制的探索。

关键词　公平规范执行; 群体偏见; 纯粹偏好理论; 规范聚焦理论

1 引言

古往今来，公平都是人类追求的永恒目标之一。论语中有“不患贫而患不均”的思想，而公平正义也是当代社会主义核心价值观的重要内容。公平规范执行(fairness norm enforcement)是人们自愿损耗利益来惩罚违反公平原则行为的意愿和行为(Feng, Luo, & Krueger, 2015)。国内外学者大多借助于反映真实生活特征的资产分配任务，如最后通牒博弈(ultimatum game, UG)、第三方最后通牒博弈(third party ultimatum game, TPUG)和第三方惩罚博弈(third party punishment game, TPPG)，来描述与探究个体的公平规范执行行为(杨邵峰, 齐春辉, 张志超, 张振, 2018; Biella & Sacchi, 2018; Everett, Faber, Crockett, & De Dreu, 2015; Jordan, McAuliffe, & Warnekena, 2014; Reimersa, Büchelb, & Diekhofa, 2017; Wu & Gao, 2018; Yudkin, Rothmund, Twardawski, Thalla, & Van Bavel, 2016)。大量研究显示人们在资产分配情景中存在极强的公平偏好(Zheng, Yang, Jin, Qi, & Liu, 2017)，对不公的分配方案表现出强烈的厌恶性(Wang, Li, Li, Wei, & Li, 2016)，并且往往会表现出公平规范执行(Henrich et al., 2006)。公平规范执行被视为一种违背理性人假设的社会偏好行为(张慧, 马红宇, 徐富明, 刘燕君, 史燕伟, 2018)，更适合从认知、情绪和动机角度进行解释(罗艺, 封春亮, 古若雷, 吴婷婷, 罗跃嘉, 2013)，同时容易受到人格因素和环境因素的影响(Güth & Kocher, 2014)。

[1]收稿日期：2019-03-22
* 基金项目：河南省哲学社会科学规划项目（2019CJY030）、河南省软科学研究计划项目（192400410021）、国家自然科学基金项目(31860280)、河南省教育厅人文社会科学研究项目（2020-ZDJH-152）、河南师范大学博士科研启动基金（QD18043）和河南师范大学青年科学基金（2019QK31）资助。
通讯作者：齐春辉, E-mail: qchizz@126.com; 高晓雷, E-mail: gaoxiaolei2010@163.com

人类是群居生活的社会性物种，群体协作与互动贯穿于整个社会生活，也是人类得以生存、发展和繁衍的关键因素。群体认同(group identity)是个体认可自己所属某个社会群体的身份及其所付诸于该群体身份上的价值与情绪(Cikara & Van Bavel, 2014; Everett, Faber, & Crockett, 2015; Tajfel & Turner, 1979)。随着当今社会全球通信、外交、经济贸易和文化交流的快速发展与日益频繁，不同种族、社会或文化群体成员之间开展合作与协作的必要性和重要性亦日趋凸显。组织、社区和社会所面临的一个巨大挑战是人们常常会区分出内群体和外群体(Hewstone, Rubin, & Willis, 2002)。当群体身份被划分并得到认同后，会促使个体表现出群体偏见(intergroup bias)或狭隘主义(parochialism)，即对内群体更为善意、宽容和利他，而对外群体则更为猜忌、漠视甚至敌意(Brewer, 1999; Hewstone et al., 2002)。例如，相比于外群体成员，个体与内群体成员存在更多的合作行为(Balliet, Wu, & De Dreu, 2014)，对内群体成员的不公平行为也给予更弱的惩罚(McAuliffe & Dunham, 2016)。

近年来，群体认同对公平规范执行的影响已成为该领域的热门话题(Biella & Sacchi, 2018; McAuliffe & Dunham, 2017; Morese et al., 2016; Reimers et al., 2017; Wang et al., 2017; Wu & Gao, 2018; Yudkin et al., 2016)，得到诸多学科领域内研究者的关注，诸如社会学(Dorrough & Glöckner, 2016)、进化科学(Delton & Krasnowd, 2017)、心理学(McAuliffe & Dunham, 2016)、神经科学(Reimersa et al., 2017)等，并取得了比较丰硕的研究成果。学者们普遍赞同公平规范执行中群体偏见现象的存在，但对于其效应方向性、作用机制及边界范畴仍存在分歧和争论：大多数学者发现群体认同会导致内群体偏爱，进而削弱针对内群体成员的公平规范执行(McAuliffe & Dunham, 2016)，而少数学者则发现内群体成员的公平违背会导致“黑羊效应”(the black sheep effect, BSE)，进而增强公平规范执行(Abrams, Palmer, Rutland, Cameron, & Van de Vyver, 2014)。对以往研究成果的系统梳理不仅有益于人们对此现象的把握与理解，澄清理论观点之间的分歧，而且能够为理论争议的解决提供助力。因此，本文从多视角总结了该议题的实证数据，比较了目前两种解释公平规范执行中群体偏见的理论模型，并据此提出未来研究的发展方向。

2 群体认同导致公平规范执行中群体偏见的实证数据

以“群体认同影响公平规范执行”为议题的实证研究往往集中于行为经济学、社会心理学和神经科学领域，大多以行为博弈研究为主，脑成像研究为辅，但却比较零散。本研究旨在系统梳理先前的研究数据，以一种相互拮抗的竞争性框架，更为清晰地呈现群体认同影响公平规范执行的方向和形式。换言之，群体认同到底是妨碍公平规范执行，还是促进公平规范执行？

2.1 内群体偏爱：群体认同妨碍公平规范执行

大多数研究发现群体认同会妨碍公平规范执行，表现出内群体偏爱效应。社会群体可以依赖于大范围的主观或客观标准进行划分，主要包括虚拟线索、自然线索和社会线索(佐斌, 温芳芳, 宋静静, 代涛涛, 2019; 温芳芳, 佐斌, 2018)。梳理总结先前的研究，基于虚拟线索、自然线索和社会线索所形成的群体认同都会削弱公平规范执行(见图 1)。

2.1.1 虚拟线索

最简单群体范式(minimal group paradigm, MGP)是探究群体关系最有影响力的实验范式(Otten, 2016; Tajfel et al., 1971)，通过虚假反馈或随机分组等操纵形成具有社会心理意义的内群体与外群体(Lane, 2016)。目前，MGP 已发展成为包括经典任务程序、随机分配程序、想象程序和计算机模拟等形式的实验范式(温芳芳, 佐斌, 2018)，且系列研究发现虚拟线索形成的群体认同能够妨碍成人和儿童对内群体成员的公平规范执行。在成人样本研究中，有学者采用 MGP 操纵互动双方的群体身份，并要求反应者分别与内/外群体成员完成 UG，结果发现内群体成员的不公平提议接受率显著高于外群体成员(王益文 等, 2014; 张瀚月, 赵玉芳, 2018; Wang et al., 2017)。

同时，Jordan 等(2014)采用 MGP 范式操纵群体认同，要求 6 或 8 岁儿童作为第三方旁观者决定是否有偿的惩罚自私分享行为，发现当外群体成员向内群体成员提供不公平提议时， 6 岁和 8 岁儿童都会给予更严厉的惩罚。

2.1.2 自然线索

自然线索主要指性别、种族、年龄等外显视觉凸显因素，具有很强的社会分类加工优势(Weisman, Johnson, & Shutts, 2015)。一些研究就采用自然线索(如种族)来操纵互动双方群体关系，发现公平规范执行中存在显著的内群体偏爱现象。徐丹妮、李建升和陈硕(2012)通过呈现不同种族提议者的面孔来实现群体分组，发现受测者对同族提议者所提不公提议的接受率显著高于异族提议者。Kubota 等采用种族信息检验了种族歧视对公平加工的影响，发现美国白人更愿意接受白人提供的分配提议，尤其是不公提议，而且受测者自身的内隐种族偏见越高，其效应越强烈，最终将上述效应归因于美国白人对黑人存在着敌意、攻击的刻板印象和偏见(Kubota, Li, Bar-David, Banaji, & Phelps, 2013)。

2.1.3 社会线索

相较于随机线索和自然线索而言，社会线索往往能够提供更多的社会化信息和意义，包括国籍、校友、兴趣爱好等(佐斌 等, 2019)。目前研究比较一致地发现多种社会线索均可有效诱发群体认同，并妨碍公平规范的执行(柏子琳, 伍海燕, 方永超, 韩红, 牛盾, 2018; 郭庆科, 徐萍, 吴睿, 胡姗姗, 2016; Apps, McKay, Azevedo, Whitehouse, & Tsakiris, 2018; Brüne et al., 2012; Feng et al., 2017; Mills, Tainsky, Green, & Leopkey, 2018; Valenzuela & Srivastava, 2012; Wang, Zheng, Meng, Lu, & Ma, 2016)。首先，国籍及其附属的语言能够诱发凸显的群体认同，进而妨碍公平规范执行例如，柏子琳等(2018)利用改编的最后通牒博弈范式，发现方言条件下不公平提议(8:2)接受率显著高于普通话条件，认为老乡情结产生了内群体偏爱、安全感和归属感，削弱了个体对绝对公平的追求。其次，鉴于一些研究往往以学生为研究样本，因此校友、玩伴等也能产生强烈的群体区分，导致公平规范执行中的内群体偏爱现象。如 Brüne 等(2012)通过告知反应者与提议者是否来自于其同伴群体，来操纵反应者与提议者之间的组内-组外区分，发现反应者更愿意接受内群体成员提供的不公平提议。郭庆科等(2016)采用隶属相同学校来操纵群体身份，要求小学一、三和五年级学生参与第三方惩罚任务，结果发现一年级学生的第三方惩罚不受群体认同的影响，但三年级和五年级学生对外群体的惩罚强度显著高于内群体，表明三年级小学生具备了明显的内群体偏好。最后，为了拓展研究结果的外部效度，新近研究者倾向于采用兴趣偏好 (如喜爱某支足球队或篮球队等)来操控群体关系，也重复验证了群体认同对公平规范执行的妨碍作用。例如，Mills 等(2018)采用双方是否喜爱相同足球队来操纵群体身份，探讨运动管理中竞技运动球迷彼此之间的行为，结果发现反应者对内群体成员的提议接受率明显高于外群体成员。Apps 等人(2018)同样采用提议者和反应者的球迷身份设置了外群体互动、内群体互动和中性互动三种条件，结果表明人们愿意消耗利益阻止外群体成员从不公平或公平提议中获益，更倾向于拒绝外群体成员的不公平提议。

2.2 黑羊效应：群体认同促进公平规范执行

另一些研究者则发现群体认同也会诱发黑羊效应，表现为人们对内群体成员的不公行为会给予更严厉惩罚和制裁。虽然此类研究较少，但是学者们仍然在虚拟线索、自然线索和社会线索范畴中，比较稳健的重复印证了群体认同对公平规范执行的提升效应。

2.2.1 虚拟线索

新近 Wu 和 Gao(2018)以 3~6 岁中国儿童为对象，采用最简单群体范式操纵群体身份，检验了不同发展阶段的儿童如何处理内群体偏爱与公平规范执行之间的冲突，结果发现 5~6 岁女童更倾向于惩罚内群体违规者，表明存在一种维持群体公平规范的信号，而 5~6 岁男童会同等程度地惩罚内/外群体违规者，只是惩罚外群体成员的强度要高于女童，这种结果部分程度上支持

了黑羊效应。

2.2.2 自然线索

有研究者通过性别和种族等自然线索操纵群体关系，发现公平规范执行中存在黑羊效应。王芹和白学军(2010)在成人群体中进行博弈双方性别配对，发现男性反应者更愿意接受女性提议者给出的不公平提议，更多拒绝男性提议者的不公平提议；而女性反应者则不会受到对方性别的影响。新近王芹等人在儿童样本中同样发现男性配对互动时存在更高的拒绝率(王芹, 白学军, 袁心颖, 尹吉端, 2018)。国外学者采用种族配对来操纵内群体互动与外群体互动，发现被试更倾向于拒绝内群体成员提出的中等不公平提议，证实了成人被试更倾向于制裁内群体成员的公平规范违背行为(Mendoza, Lane, & Amodio, 2014)。

2.2.3 社会线索

基于同伴、学校、球队偏好等社会线索形成的群组区分，也被多名学者证实会导致黑羊效应。例如，McLeish 和 Oxoby(2011)启动互动双方是否隶属于相同同伴群体，并要求被试报告其在 UG 中的最低可接受提议，发现内群体启动下被试的最低可接受提议值要显著高于外群体启动 。Mendoza 等人(2014)通过校友身份操纵群体区分，重复印证了公平规范执行中的黑羊效应，而且拥有强烈群体认同的个体存在更明显的效应。Morese 等(2016)则采用足球球迷认同来操纵成人被试的群体关系，发现被试参与第三方惩罚任务时更愿意惩罚旨在伤害其他内群体成员的内群体成员。

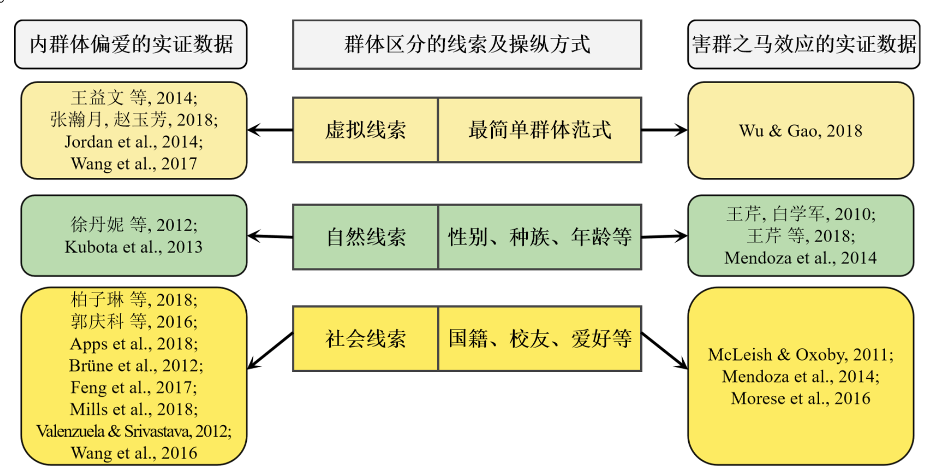


图 1 公平规范执行中群体偏见的实证数据示意图

3 群体认同导致公平规范执行中群体偏见的理论基础

当群体认同与公平规范相遭遇时，人们便会面临一种更复杂的两难情景。人们如何认知并应对内群体成员的不公平行为呢？该困境涉及到内群体偏爱动机与黑羊效应的相互权衡，也体现了人情与道德之间的抉择。目前研究者主要采用纯粹偏好理论和规范聚焦理论来解释公平规范执行中群体偏见的发生机制。

3.1 纯粹偏好理论

多数学者往往采用偏好来解释群体互动中的狭隘主义，认为人们对内群体成员有简单且强烈的亲社会偏好，诸如信任、合作、宽容等(Hogg, Abrams, & Brewer, 2017)。纯粹偏好理论(mere preference theory, MPT)试图从内群体偏爱角度来解释人类如何应对内群体成员的公平违规行为，认为当人们对其所隶属的群体形成清晰认同后，会倾向于采用积极、正性的方式评估其所属群体及成员，这种对内群体及其成员更为积极的认知与评价会使其更加溺爱、宽容、包庇内群体成员的过失，或者对于过失行为给予合理化的解释，最终使得内群体成员承受较弱的惩罚(Choi &

Bowles, 2007; McAuliffe & Dunham, 2016)。

纯粹偏好理论认为内群体积极评价和行为合理化是导致内群体偏爱现象的重要路径。一方面依据社会认同理论，人们倾向于采用正性和区分性的方式评估任何与自我有关的事物。当群体区分形成并得到认同后，人们也更愿意积极评价内群体成员(Tajfel & Turner, 1979)。这种“内群体依恋和积极性”(Brewer, 2006)驱使人们更愿意忍受内群体成员的不公行为。简言之，内群体成员的负性违规行为会被群体认同诱发的积极评价所抵消，进而有效降低被惩罚的可能性与强度(McAuliffe & Dunham, 2016)。如张瀚月和赵玉芳(2018)采用最简单群体范式操控群体身份，要求受测者作为反应者分别与内、外群体成员完成最后通牒博弈，结果发现人们对内群体成员的不公平行为接受率显著高于外群体成员，认为内群体积极评价增强了受测者的公平耐受性。柏子琳等(2018)利用改编的最后通牒博弈范式，发现方言条件下不公平提议(8:2)接受率显著高于普通话条件，认为老乡情结产生了内群体偏爱、安全感和归属感，削弱了个体对绝对公平的追求。

另一方面，行为合理化是指人们对内群体成员的不公平行为潜在的动机、意图和目的等进行合理解释和归因，主要涉及到心理理论加工，即指个体了解自己和他人愿望、信念、意图等心理状态并据此推测他人行为的能力(王益文, 张文新, 2002)。有研究表明对内群体成员公平提议的接纳涉及到负责心理推理功能的大脑网络的激活(Baumgartner, Götte, Gügler, & Fehr, 2012; Fatfouta, Meshi, Merkl, & Heekeren, 2018)。Baumgartner 等人(2012)及其同事发现个体对内群体成员的自私行为给予宽大处理，惩罚强度与心理推理网络激活程度负相关，好像其试图理解并合理化内群体成员的过失行为。Fatfouta 及其同事(2016)发现人们更愿意接纳恋人提供的不公平提议，且接纳恋人的不公平提议会增强内侧前额叶皮质(medial prefrontal cortex, mPFC)的激活及其与背侧前扣带回(dorsal anterior cingulate cortex, dACC)之间的功能连接，其中 mPFC 隶属于心理理论网络，反映了人们对不公行为的解释和合理化；dACC 负责冲突信号加工，mPFC 与 dACC 之间功能连接的增强也反映了人们借助合理化加工来解决认知冲突的过程。

作为最常用来解释公平规范执行中群体偏见效应的理论，纯粹偏好理论具有非常突出的优势。首先，纯粹偏好理论能够合理地阐明群体认同对公平规范执行的妨碍作用，表明该理论具有很强的解释力和适用性(Balliet et al., 2014; McAuliffe & Dunham, 2016)。其次，内群体积极评价和行为合理化的作用路径侧重于个体内在的动机因素，如自我服务动机等，这是一种近端解释机制(刘长江, 郝芳, 2014)，相对简单易懂，容易被大多数学者所接纳。最后，内群体偏爱反映了群体认同所驱动的一种群际行为模式，此时个体的行为动机在一定程度上是服务于自我利益的，这种功利性动机为解释群体合作行为提供了一种良好的思路。然而，纯粹偏好理论也存在一些不足。一方面，群体互动是一种牵涉不同群体身份他人的交互式情境，必然包含背叛或欺骗的风险和不确定性，纯粹偏好理论仅聚焦于人们的内群体偏爱，忽略了决策主体对互动他人及情境的主观风险评估的影响(Valenzuela & Srivastava, 2012)。另一方面，纯粹偏好理论没有关注群体规范的重要性，因此无法解释黑羊效应，即当内群体成员违反某些群体规范时会遭受严厉的制裁，诸如驱逐、流放等(Wang, Zheng, Meng, Lu, Ma, 2016)。

3.2 规范聚焦理论

一些学者则强调规范在群体形成、运作和维持中所起的重要作用，认为人类的群体生活就是为了解决协作问题，即合作问题(Balliet et al., 2014)。群体通过传授、教导内群体成员如何行为的规章制度，就能够提升内群体成员的可预测性，进而促进了群体互动的顺畅与便捷(McAuliffe & Dunham, 2016)。而且，群体成员通过谨慎的遵循并内化规范，获得与其他成员良性互动的经验和利益，会进一步强化群体规范的有效性、适用性和认可性(Bowles, 2006)。基于此，规范聚焦理论(normal focus theory, NFT)试图从内群体规范视角来理解人类如何应对内群体成员的公平违规行为，强调个体对内群体的依附是功能性的，能够强化个体对群体规范的遵循和合作预期，而且群体也是一种养育合作行为的社会结构，内群体成员的违规行为违反了群体生活的核心原

则，可视为群体认同的一种潜在威胁，因此人们会严厉地惩罚内群体成员的不公平行为(McAuliffe & Dunham, 2016)。

规范聚焦理论认为黑羊效应主要是由预期违背认知和规范维持动机两方面因素导致的。预期违背主要体现在个体水平上，反映了人们对实际结果偏离事先预期的认知。预期违背理论(Brewer, 1999; Hewig et al., 2011)认为，鉴于内群体成员彼此分享着相似的价值和规范，人们往往预期内群体成员具备更高的合作性和值得信赖性，因此内群体成员的不公行为严重偏离了人们的事先预期，使人产生更强烈的负性情绪(如愤怒、厌恶等)，进而导致更严厉的制裁。Mendoza等人(2014)发现，相比于外群体成员，内群体成员的不公行为会遭到更严厉的制裁；群体认同与此效应存在显著正相关，并能增强受测者对内群体成员的互惠预期。脑电研究也证实了内群体成员的不公行为会导致更强烈的预期违背(王益文 等, 2014; Wang et al., 2017; Wu, Hu, Van Dijk, Leliveld, & Zhou, 2011)。王益文及其团队采用 UG 任务和 MGP 范式，探讨了反应者如何评估来自内群体或外群体成员的公平或不公平提议，结果发现内群体成员的不公平提议比公平提议诱发更强烈的反馈相关负波(feedback-related negativity, FRN)，而外群体互动时却未发现提议的显著效应(王益文 等, 2014; Wang et al., 2017)。鉴于 FRN 负责编码反馈结果与预期之间的偏差(即预期奖赏错误信号)，预期奖赏错误信号越大，FRN 波幅越大，因此该结果表明内群体互动时存在强烈的互惠预期，不公平提议导致了极大的预期违背。

相较而言，规范维持动机是一种体现在群体水平上的宏观机制，反映了人们拥护和维持其所属群体核心规范的需求。群体规范维持理论认为，群体规范是形成、运作和维持群体的必要条件，也是群体成员所认可、遵循并内化的行为准则，人们为了维持群体规范会对自私的内群体成员给予严厉的制裁(Bernhard, Fischbacher, & Fehr, 2006; Shinada, Yamagishi, & Ohmura, 2004)。换言之，当内群体成员的行为严重偏离群体规范时，往往被视为一种有损群体规范或声誉的潜在威胁，为了有效应对其潜在损害并防止类似事件再次发生，人们也会对这种偏离行为给以严厉惩罚。例如，Shinada 及其同事(2004)发现人们更愿意惩罚不合作的内群体成员而非不合作的外群体成员，认为这种针对内群体成员的制裁也是为了维持群体规范和正常运转。新近 McLeish 和 Oxoby(2011)采用启动方法操纵群体关系，要求受测者指明各种群体互动条件下其参与 UG 任务中的最低可接受提议，发现内群体启动下最低可接受提议值显著高于外群体启动，表明内群体规范违背被知觉为一种认同威胁，人们愿意给以强烈的惩罚来应对此威胁。

规范聚焦理论的优势体现在三个层面。第一，规范聚焦理论尤为突出群体规范对个体行为表现的指导、约束和矫正作用，能够较好的解释黑羊效应，认为内群体规范的违背会导致强烈的预期违背，且被知觉为一种群际威胁，内群体成员为了维持和保护群体的凝聚力与规范，会严厉惩罚其他内群体成员的不公行为(Wang et al., 2016)。第二，规范聚焦理论强调内群体规范及其违背对群体维持与运作的重要性，将群际互动中互惠预期、威胁感知等概念纳入群际合作的动力范畴，提供了一种中端水平的理论解释(Halevy & Katz, 2013)。第三，与纯粹偏好理论所强调的内群体偏爱动机相反，规范聚焦理论所提倡的内群体规范是一种非功利性动机，旨在维持与保护内群体凝聚力和协作，拓展了群体合作行为产生缘由的范畴(Delton & Krasnow, 2017)。虽然，规范聚焦理论能够解释黑羊效应等，但其仍有一些不足。一方面，现有研究大多均支持纯粹偏好理论的预期，而规范聚焦理论的支撑证据仅存在于少量研究，在一定程度上表明规范聚焦理论的适应性和有效性相对较差。另一方面，与纯粹偏好理论过于强调内群体偏爱动机相似，聚焦内群体规范的规范聚焦理论也存在忽视决策主体社会偏好的特点，如社会价值取向、不平等厌恶等(Everett et al., 2015)。

4 展望

由上述文献回顾可知，近年来国内外学者对群体认同影响公平规范执行展开了日益系统的

探究，比较稳健地证实了公平规范执行中的群体偏见现象。尽管取得了比较丰硕的成果，但该领域仍存在一些尚未解决的议题，例如此效应的发生机制和作用边界又如何？多种线索操纵所致偏见是否一致？内群体偏爱和黑羊效应能否进行整合？公平规范执行中群体偏见潜在的神经机制是什么？这些都有待于未来研究的深入有序检验。

4.1 揭示公平规范执行偏见的边界条件

现有成果的整理与分析表明群体认同会影响公平规范执行，但是这些数据仍未清晰详细的阐明公平规范执行中狭隘利他性的边界条件。首先，群体偏见往往包含内群体偏爱和外群体贬损两种条件，虽然现有大多数研究倾向于聚焦内群体偏爱，但是却不能完全否认外群体贬损的作用。然而，现有研究大多仅仅设置内群体成员或外群体成员两种潜在互动条件，这种缺乏中性基线的二分设置无法有效区分内群体偏爱和外群体贬损影响公平规范执行的相对权重(Schiller, Baumgartner, & Knoch, 2014)。例如，新近 Apps 等(2018)要求被试作为反应者，分别与支持同一支足球队的个体(内群体成员)、支持主要敌对球队的个体(外群体成员)和中性个体共同完成 UG 任务，结果发现人们愿意消耗利益阻止外群体成员从不公平或公平提议中获益，更倾向于拒绝外群体成员的不公平提议。因此，未来研究应该通过设置一个中性控制(即设置内群体成员、组外成员和未分类成员)，进而清晰地回答这个问题。

其次，公平规范执行可以通过第二方任务和第三方任务进行操纵与测量。第二方任务往往是 UG 博弈，要求被试直接参与互动并协调自我利益与公平规范执行之间的关系。第三方任务多是 TPUG 和 TPPG，要求被试作为旁观者，对其他两人之间的分配不公现象进行干预。例如，Jordan 等(2014)采用 MGP 范式操纵群体认同，要求 6 或 8 岁儿童作为第三方旁观者决定是否有偿的惩罚自私分享行为，发现当外群体成员向内群体成员提供不公平提议时，6 岁儿童会给予更严厉的惩罚；8 岁儿童同样会严厉惩罚自私的外群体成员，但也会惩罚不利于内群体或外群体成员的不公平行为，表明规范执行从其出现时就受到群体身份的调控，但这种内群体偏爱会随着年龄发展而受到削弱。由于两种任务在决策视角、收益矩阵、度量尺度等上均存在显著的差异，且现有研究大多聚焦于第二方的最后通牒博弈任务，因此未来研究仍需加强第三方视角下公平规范执行的狭隘利他性研究，揭示第二方和第三方视角下该效应的一致性与差异性，确定其效应边界。

最后，现有研究多以成人为被试，发现内群体偏爱与公平规范执行之间存在很强的张力，导致上述彼此混淆的结果。考虑到成人被试具有很丰富的群体互动经验和社会文化熏陶，因此从个体心理发展视角出发，检验儿童如何权衡内群体偏爱动机和公平规范执行动机之间的竞争需求，在一定程度上能够将人类的早期倾向与晚期社会化结果相分离，有助于理解公平规范执行潜在的发生机制以及人类群体中心性的特点(Wu & Gao, 2018; McAuliffe & Dunham, 2017)。新近 McAuliffe 和 Dunham(2017)采用 MGP 范式来操纵 6 至 10 岁儿童的群体认同，并在没有欺骗的情境中检验了儿童与内/外群体成员的谈判行为，结果发现 MGP 能够有效的诱发内群体偏见，却无法调控儿童对分配提议的反应，表明儿童作为违规行为的直接受害者时，更加偏爱公平规范执行而非内群体偏爱。

4.2 比较多种线索操纵所致偏见的差异性

如前所述，社会群体的操纵标准可以区分为虚拟线索、自然线索和社会线索三种(佐斌 等, 2019; 温芳芳, 佐斌, 2018)，但不同操纵所形成的群体概念的多样性和差异性也可能是导致现有结果变异性较大的重要缘由(严磊, 佐斌, 张艳红, 吴漾, 杨林川, 2018)。大量证据表明三种线索建构的群体认同是存在差异的，而且会诱发不同的群体偏见效应。首先，有研究发现最简单群体范式中被试给予内群体违规者的惩罚要小于其给予组外违规者的惩罚强度；而在真实群体当中，被试给予内群体违规者的惩罚强度要远高于组外违规者(Goette, Huffman, & Meier, 2006)。其次，

虽然学者们认为真实群体身份要比最简单群体身份具有更强的群体认同(Huettel & Kranton, 2012)，但是一项元分析文献表明歧视依赖于所研究的群体认同，表现为随机线索诱导的群体认同要比依据种族、国籍或其他社会属性所产生的群体认同更为有效(Lane, 2016)。最后，很多社会线索，诸如老乡、兴趣偏爱等，往往受到社会化过程、地缘位置等因素的影响，具有极强的文化差异性。例如，基于社会线索建构群体关系进而发现黑羊效应的研究大多发现于西方文化背景，可能就反映了东方文化更侧重于集体主义文化，对内群体成员的过失表现出更多的忍让和包容(Wang et al., 2017)。因此，未来研究尚需仔细思考不同群体认同的本质特性如何影响规范执行，公平规范执行是否特异地存在于某种内群体标准之中。换言之，某些群体可能具有强烈的内群体慷慨与合作规范，此时内群体成员的违规行为将会导致更严厉的惩罚，而其他群体则不存在这种效应。

4.3 促进理论观点的融合与互补

鉴于群际公平规范执行的复杂性，人们试图从内群体偏爱和规范维持来理解其产生机制(Delton & Krasnow, 2017)。虽然两者目前对群际公平规范执行的方向性和作用机制尚存分歧，但彼此之间仍存在非常紧密的关联。一方面，有学者认为内群体偏爱和黑羊效应在最终目标上也是相一致的(Mendoza et al., 2014; Ellemers & Jetten, 2013)。Mendoza 等(2014)发现人们更倾向于拒绝内群体成员的不公提议，将此视为一种更高层次的维持和提升内群体偏爱的重要策略，即个体通过严厉地惩罚内群体成员的不公行为，进而巩固群体价值和维持群体凝聚力，最终维持内群体偏爱。这种解释也与现有研究结果相一致，即内群体成员对规范性合作义务的感知会直接决定预期惩罚和实际惩罚的强度，合作责任的感知性越强，内群体惩罚强度越强(Goette et al., 2006; Valenzuela & Srivastava, 2012)。另一方面，Ellemer 和 Jetten(2013)强调群体成员可以区分为核心成员和边缘成员，群体内的边缘状态存在多种作用路径，是个体与群体协商融合的产物。该理论认为核心成员自认为拥有惩罚那些达不到群体要求或违反群体规范成员的责任和义务，这种制裁和惩罚会强化集体意识的理解，促进群体边界的划分，有益于阐明那些指导群体行为的规范(Doosje, Spears, Ellemers, & Koomen, 1999)。因此，黑羊效应看似与内群体偏爱态度相悖，但其实际上却是服务于群体的利益，有助于维持和提升内群体偏爱，未来研究应试图在此整合性的框架中探究公平规范执行中的群体偏见现象。

4.4 增强公平规范执行偏见的神经机制研究

随着认知神经科学的兴起与脑成像技术的应用，国内外学者日益尝试于揭示公平规范执行偏见的潜在神经机制，其中采用 ERP 和 fMRI 技术的脑成像研究目前处于优势地位。一方面，ERP 技术具有较高的时间分辨率，有助于准确描绘认知加工动态分离的时间进程。王益文等探讨了最简单群体互动情景中公平规范执行偏见的脑电机制，发现内群体成员的不公提议比公平提议产生更负的 FRN，而外群体互动则无此效应(王益文 等, 2014; Wang et al., 2017)。Wang 等(2016)进一步考察了敌意意图模糊性调节公平规范执行偏见的脑电机制，发现当提议者的敌意意图比较清晰时，内群体成员提供的不公提议要比外群体成员诱发更负的 FRN，为黑羊效应提供了神经证据；而当敌意意图被感知为模糊时，FRN 表现为相反的模式，为内群体偏爱提供了神经证据。因此，FRN 是检测公平规范执行中群体偏见存在何种表现形式的重要参考指标。

另一方面，fMRI 技术则具有较高的空间分辨率，有利于评估认知加工潜在的脑区激活模式及各脑区间的功能连接。Morese 等(2016)检验了第三方惩罚情景中公平规范执行偏见的潜在神经基础，结果发现，相较于外群体成员，内群体成员的不公行为会显著增强 mPFC 和 TPJ 的激活程度，表明旁观者试图理解或合理化内群体成员的违规行为。Apps 等(2018)则在 UG 任务中再次证实了上述结果，结果发现，内群体互动时不公提议比公平提议诱发更强烈的 mPFC 激活程度，然而外群体或中性群体互动时则不存在类似模式；同时，受测者对内群体成员的认同程度越高，

此交互作用的效应越强。新近 Fatfouta 等(2016)进一步证实，亲密他人所提的不公提议诱发了更强烈的 mPFC 激活以及 mPFC 与 dACC 间的功能连接；同时，那些存在较弱 mPFC-dACC 的功能连接的被试更倾向于接受亲密恋人的不公提议。因此，负责合理化他人意图的心理理论网络（包括 mPFC 和 TPJ）和负责冲突信号检测加工的认知控制网络（如 dACC）均参与到公平规范执行的群体偏见现象当中。

然而，当前研究大多聚焦于使用脑成像方法来揭示参与认知加工的特定脑区或脑电成分，未来研究应侧重于采用无创脑刺激技术考察特定脑区与认知加工的因果关联。rTMS 和 tDCS 能够有针对性的兴奋或抑制特定脑区的认知功能，更适合作出特定脑区所起功能的因果性论断，因此未来研究应多借助此两种技术开展研究(董军, 付淑英, 卢山, 杨绍峰, 齐春辉, 2017)。另外，随着计算机科学方法的发展，未来研究也应考虑脑成像数据与计算模拟方法的结合，尤其是利用多种脑成像数据指标进行模型建构。Friston 等(2014)发展了一种计算模拟框架，用于理解社会学习中奖赏预期错误如何调控个体的行为选择，因此脑科学与计算科学的交叉研究是一种值得期待的研究取向。

参考文献

柏子琳, 伍海燕, 方永超, 韩红, 牛盾. (2018). 方言对社会决策及情绪的影响——来自电生理的证据. *心理科学, 41*(5), 1171–1177.

董军, 付淑英, 卢山, 杨绍峰, 齐春辉. (2017). 自我控制失败的理论模型与神经基础. *心理科学进展, 26*(1), 134–143.

郭庆科, 徐萍, 吴睿, 胡姗姗. (2016). 群体偏好与年级对小学生利他惩罚行为的影响. *心理发展与教育, 32*(4), 402–408.

刘长江, 郝芳. (2014). 社会困境问题的理论架构与实验研究. *心理科学进展, 22*(9), 1475–1484.

罗艺, 封春亮, 古若雷, 吴婷婷, 罗跃嘉. (2013). 社会决策中的公平准则及其神经机制. *心理科学进展, 21*(2), 300–308.

王芹, 白学军, 袁心颖, 尹吉端. (2018). 经济博弈中不同性别儿童的“以貌取人”对决策行为的影响. *内蒙古师范大学学报(自然科学汉文版), 47*(6), 522–526.

王芹, 白学军. (2010). 最后通牒博弈中回应者的情绪唤醒和决策行为研究. *心理科学, 33*(4), 844–847.

王益文, 张文新. (2002). 3~ 6 岁儿童“心理理论”的发展. *心理发展与教育, 18*(1), 11–15.

王益文, 张振, 张蔚, 黄亮, 郭丰波, 原胜. (2014). 群体身份调节最后通牒博弈的公平关注. *心理学报, 46*(12), 1850–1859.

温芳芳, 佐斌. (2018). 最简群体范式的操作, 心理机制及新应用. *心理科学, 41*(3), 713–719.

徐丹妮, 李建升, 陈硕. (2012). *社会分组影响对不公平分配的决策反应.* 第十五届全国心理学学术会议论文摘要集.

严磊, 佐斌, 张艳红, 吴漾, 杨林川. (2018). 交叉分类及其对刻板印象的影响. *心理科学进展, 26*(7), 1272–1283.

杨邵峰, 齐春辉, 张志超, 张振. (2018). 价值取向对自我他人决策时公平规范执行的影响. *心理与行为研究, 16*(6), 834–840.

张瀚月, 赵玉芳. (2018). 社会距离对不公平行为回应的影响. *西南大学学报 (自然科学版), 40*(2), 140–145.

张慧, 马红宇, 徐富明, 刘燕君, 史燕伟. (2017). 最后通牒博弈中的公平偏好: 基于双系统理论的视角. *心理科学进展, 26*(2), 319–330.

佐斌, 温芳芳, 宋静静, 代涛涛. (2019). 社会分类的特性, 维度及心理效应. *心理科学进展, 27*(1), 141–148.

Abrams, D., Palmer, S. B., Rutland, A., Cameron, L., & Van de Vyver, J. (2014). Evaluations of and reasoning about normative and deviant ingroup and outgroup members: Development of the black sheep effect. *Developmental Psychology, 50*(1), 258–270.

Apps, M., Mckay, R., Azevedo, R. T., Whitehouse, H., & Tsakiris, M. (2018). Not on my team: Medial prefrontal cortex responses to ingroup fusion and unfair monetary divisions. *Brain & Behavior, 8*(8), e01030.

Balliet, D., Wu, J., & De Dreu, C. K. (2014). Ingroup favoritism in cooperation: A meta-analysis. *Psychological Bulletin, 140*(6), 1556–1581.

Baumgartner, T., Götte, L., Gügler, R., & Fehr, E. (2012). The mentalizing network orchestrates the impact of parochial altruism on social norm enforcement. *Human Brain Mapping, 33*(6), 1452–1469.

Bernhard, H., Fischbacher, U., & Fehr, E. (2006). Parochial altruism in humans. *Nature, 442*(7105), 912–915.

Biella, M., & Sacchi, S. (2018). Not fair but acceptable… for us! Group membership influences the trade off between equality and utility in a Third Party Ultimatum Game. *Journal of Experimental Social Psychology, 77*, 117–131.

Bowles, S. (2006). Group competition, reproductive leveling, and the evolution of human altruism. *Science, 314*(5805), 1569–1572.

Brewer, M. B. (1999). The psychology of prejudice: Ingroup love and outgroup hate?. *Journal of Social Issues, 55*(3), 429–444.

Brüne, M., Tas, C., Wischniewski, J., Welpinghus, A., Heinisch, C., & Newen, A. (2012). Hypnotic ingroup–outgroup suggestion influences economic decision-making in an ultimatum game. *Consciousness & Cognition, 21*(2), 939–946.

Choi, J. K., & Bowles, S. (2007). The coevolution of parochial altruism and war. *Science, 318*(5850), 636–640.

Cikara, M., & Van Bavel, J. J. (2014). The neuroscience of intergroup relations: An integrative review. *Perspectives on Psychological Science, 9*(3), 245–274.

Delton, A. W., & Krasnow, M. M. (2017). The psychology of deterrence explains why group membership matters for third-party punishment. *Evolution and Human Behavior, 38*(6), 734–743.

Doosje, B., Spears, R., Ellemers, N., & Koomen, W. (1999). Perceived group variability in intergroup relations: The distinctive role of social identity. *European Review of Social Psychology, 10*(1), 41–74.

Dorrough, A. R., & Glöckner, A. (2016). Multinational investigation of cross-societal cooperation. *Proceedings of the National Academy of Sciences, 113*(39), 10836–10841.

Ellemers, N., & Jetten, J. (2013). The many ways to be marginal in a group. *Personality and Social Psychology Review, 17*(1), 3–21.

Everett, J. A., Faber, N. S., & Crockett, M. (2015). Preferences and beliefs in ingroup favoritism. *Frontiers in Behavioral Neuroscience, 9*, 15.

Everett, J. A., Faber, N. S., Crockett, M. J., & De Dreu, C. K. (2015). Economic games and social neuroscience methods can help elucidate the psychology of parochial altruism. *Frontiers in Psychology, 6*, 861.

Fatfouta, R., Meshi, D., Merkl, A., & Heekeren, H. R. (2018). Accepting unfairness by a significant other is associated with reduced connectivity between medial prefrontal and dorsal anterior cingulate cortex. *Social Neuroscience, 13*(1), 61–73.

Feng, C., Luo, Y. J., & Krueger, F. (2015). Neural signatures of fairness-related normative decision making in the ultimatum game: A coordinate-based meta-analysis. *Human Brain Mapping, 36*(2), 591–602.

Friston, K., Schwartenbeck, P., FitzGerald, T., Moutoussis, M., Behrens, T., & Dolan, R. J. (2014). The anatomy of choice: Dopamine and decision-making. *Philosophical Transactions of the Royal Society of London: Series B. Biological Sciences, 369*, 20130481.

Goette, L., Huffman, D., & Meier, S. (2006). The impact of group membership on cooperation and norm enforcement: Evidence using random assignment to real social groups. *American Economic Review, 96*(2), 212–216.

Güth, W., & Kocher, M. G. (2014). More than thirty years of ultimatum bargaining experiments: Motives, variations, and a survey of the recent literature. *Journal of Economic Behavior & Organization, 108*, 396–409.

Halevy, N., & Katz, J. J. (2013). Conflict templates: Thinking through interdependence. *Current Directions in Psychological Science, 22*(3), 217–224.

Henrich, J., McElreath, R., Barr, A., Ensminger, J., Barrett, C., Bolyanatz, A., ... & Lesorogol, C. (2006). Costly punishment across human societies. *Science, 312*(5781), 1767–1770.

Hewig, J., Kretschmer, N., Trippe, R. H., Hecht, H., Coles, M. G., Holroyd, C. B., & Miltner, W. H. (2011). Why humans deviate from rational choice. *Psychophysiology, 48*(4), 507–514.

Hewstone, M., Rubin, M., & Willis, H. (2002). Intergroup bias. *Annual Review of Psychology, 53*(1), 575–604.

Hogg, M. A., Abrams, D., & Brewer, M. B. (2017). Social identity: The role of self in group processes and intergroup relations. *Group Processes & Intergroup Relations, 20*(5), 570–581.

Huettel, S. A., & Kranton, R. E. (2012). Identity economics and the brain: Uncovering the mechanisms of social

conflict. *Philosophical Transactions of the Royal Society B: Biological Sciences, 367*(1589), 680–691.

Jordan, J. J., McAuliffe, K., & Warneken, F. (2014). Development of in-group favoritism in children's third-party punishment of selfishness. *Proceedings of the National Academy of Sciences, 111*(35), 12710–12715.

Kubota, J. T., Li, J., Bar-David, E., Banaji, M. R., & Phelps, E. A. (2013). The price of racial bias: Intergroup negotiations in the ultimatum game. *Psychological Science, 24*, 2498–2504.

Lane, T. (2016). Discrimination in the laboratory: A meta-analysis of economics experiments. *European Economic Review, 90*, 375–402.

McAuliffe, K., & Dunham, Y. (2016). Group bias in cooperative norm enforcement. *Philosophical Transactions of the Royal Society of London. Series B, Biological Sciences, 371*(1686), 20150073.

McAuliffe, K., & Dunham, Y. (2017). Fairness overrides group bias in children's second-party punishment. *Journal of Experimental Psychology: General, 146*(4), 485–494.

McLeish, K. N., & Oxoby, R. J. (2011). Social interactions and the salience of social identity. *Journal of Economic Psychology, 32*(1), 172–178.

Mendoza, S. A., Lane, S. P., & Amodio, D. M. (2014). For members only: Ingroup punishment of fairness norm violations in the ultimatum game. *Social Psychological and Personality Science, 5*(6), 662–670.

Mills, B. M. , Tainsky, S. , Green, B. C. , & Leopkey, B. (2017). The ultimatum game in the college football rivalry context. *Journal of Sport Management, 32*(1), 1–13.

Morese, R., Rabellino, D., Sambataro, F., Perussia, F., Valentini, M. C., Bara, B. G., & Bosco, F. M. (2016). Group membership modulates the neural circuitry underlying third party punishment. *PloS One, 11*(11), e0166357.

Otten, S. (2016). The Minimal Group Paradigm and its maximal impact in research on social categorization. *Current Opinion in Psychology, 11*, 85–89.

Reimers, L., Büchel, C., & Diekhof, E. K. (2017). Neural substrates of male parochial altruism are modulated by testosterone and behavioral strategy. *NeuroImage, 156*, 265–276.

Schiller, B., Baumgartner, T., & Knoch, D. (2014). Intergroup bias in third-party punishment stems from both ingroup favoritism and outgroup discrimination. *Evolution and Human Behavior, 35*(3), 169–175.

Shinada, M., Yamagishi, T., & Ohmura, Y. (2004). False friends are worse than bitter enemies: "Altruistic" punishment of in-group members. *Evolution and Human Behavior, 25*(6), 379–393.

Tajfel, H., & Turner, J. C. (1979). An integrative theory of intergroup conflict. *The Social Psychology of Intergroup Relations, 33*(47), 94–109.

Valenzuela, A., & Srivastava, J. (2012). Role of information asymmetry and situational salience in reducing intergroup bias: The case of ultimatum games. *Personality and Social Psychology Bulletin, 38*(12), 1671–1683.

Wang, G., Li, J., Li, Z., Wei, M., & Li, S. (2016). Medial frontal negativity reflects advantageous inequality aversion of proposers in the ultimatum game: An ERP study. *Brain Research, 1639*, 38–46.

Wang, L., Zheng, J., Meng, L., Lu, Q., & Ma, Q. (2016). Ingroup favoritism or the black sheep effect: Perceived intentions modulate subjective responses to aggressive interactions. *Neuroscience Research, 108*, 46–54.

Wang, Y., Zhang, Z., Bai, L., Lin, C., Osinsky, R., & Hewig, J. (2017). Ingroup/outgroup membership modulates fairness consideration: Neural signatures from ERPs and EEG oscillations. *Scientific Reports, 7*, 39827.

Weisman, K., Johnson, M. V., & Shutts, K. (2015). Young children's automatic encoding of social categories. *Developmental Science, 18*(6), 1036–1043.

Wu, Y., Hu, J., van Dijk, E., Leliveld, M. C., & Zhou, X. (2012). Brain activity in fairness consideration during asset distribution: Does the initial ownership play a role?. *PloS One, 7*(6), e39627.

Wu, Z., & Gao, X. (2018). Preschoolers' group bias in punishing selfishness in the Ultimatum Game. *Journal of Experimental Child Psychology, 166*, 280–292.

Yudkin, D. A., Rothmund, T., Twardawski, M., Thalla, N., & Van Bavel, J. J. (2016). Reflexive intergroup bias in third-party

punishment. *Journal of Experimental Psychology: General, 145*(11), 1448–1459.

Zheng, Y., Yang, Z., Jin, C., Qi, Y., & Liu, X. (2017). The influence of emotion on fairness-related decision making: A critical review of theories and evidence. *Frontiers in Psychology, 8*, 1592.

In-group favoritism or the black sheep effect? Group bias of fairness norm enforcement during economic games

ZHANG Zhen[1]; QI Chunhui[1]; WANG Yang[2]; ZHAO hui[1]; WANG Xiaoxin[1]; GAO Xiaolei[3]

([1] *Institute of Psychology, Henan Normal University, Xinxiang* 453007, *China*)

([2] *State Grid Tianjin Power Corporation Dongli Supply Company, Tianjin* 300300, *China*)

([3] *Educational College of Tibet University, Lhasa* 850000, *China*)

Abstract: Fairness norm enforcement refers to the willingness to incur personal costs to punish violations of fairness norms, which was thought to be a hallmark of human society and play a key role in cooperative interactions. Group identity refers to some knowledge of one's group membership together with the value and emotional significance attached to that membership, which directly influences people's fairness norm enforcement during inter-group context. Using a variety of asset allocation game, researchers found group bias exerted a critical effect on fairness norm enforcement, while existing in two opposite patterns. Sometimes, people were more likely to accept unfair offer from in-groups, reflecting the pattern of in-group favoritism, but sometimes people were also more likely to punish norm violations from in-group members, revealing the form of the so-called black sheep effect. Currently, norms focused theory and mere preferences theory have usually been used to explain the above contradictory phenomena. Based on this review, future research directions should explore the boundary conditions of this bias, compare the difference of this parochial altruism induced by variable group identity, emphasize the integration of different theories, and enhance the exploration of its underlying neural mechanisms.

Key words: fairness norm enforcement; group bias; mere preferences theory; norms focused theory